Tested Studies for Laboratory Teaching
Proceedings of the Association for Biology Laboratory Education
Vol. 36, Article 49, 2015

Mining the Genome of *Callosobruchus maculatus* (Bean Beetle) with a Little Help from Genetic Model Organisms

Marilee A. Ramesh

Roanoke College, Department of Biology, 221 College Ln., Salem VA 24153 USA
(**ramesh@roanoke.edu**)

The number of whole genome sequencing projects has increased rapidly as the technology has made sequencing at this level cheaper and easier to complete. The result has been a vast amount of raw data that could be used as fodder for inquiry-based student projects. A comparative genetic approach utilizes characterized genes from a model system as probes to search for similar genes in the genomes of less studied systems. I have developed a genetics laboratory exercise that mines genes involved in mating and life span in the bean beetles, *Callosobruchus maculatus* genome using previously characterized genes in *Drosophila melanogaster* as the initial probes. Students start with the primary literature to identify potential genes in *D. melanogaster* and collect those sequences from GenBank. Gene and protein sequences were used to probe the bean beetle genome for similar sequences. The identified sequence fragments were evaluated for validity as possible homologs. The approach can be applied to on-going genome sequencing projects, not requiring the project to be complete or annotated, thus providing the students an opportunity to work with raw sequence data.

This exercise was develop as part of the Bean Beetle: A Model Organism for Inquiry-based Undergraduate Laboratories (http://www.beanbeetles.org/index.html). Detailed resources for this specific exercise, Exploring the Genetic Basis for Behavior, are available at http://www.beanbeetles.org/protocols/genetics_behavior/synopsis.html. The appendix (part of the instructor's notes from the on-line source) provides a guide through the analysis with the couch potato gene as an example, providing expected outputs for each test. Be advised that this exercise makes use of public databases that are prone to change frequently and should be checked prior to running this exercise in class. While this exercise uses *D. melanogaster* as the model system and sexual selection as the process, the experiment could be adapted to other systems and/or other processes. Since the *C. maculatus* genome is a partial assembly, be aware that some sequence is missing and that it is likely that matches will be partial.

Appendix

Bioinformatics Tools for Mining *Callosobruchus maculatus* Genome

To Locate Protein Sequence for Your Gene of Interest

a. Go to http://www.ncbi.nlm.nih.gov/

b. There is a search bar at the top of the page. Change the default (All Databases) to **Protein**. Type in the name of the gene of interest followed by *Drosophila*.

c. The search will yield several results (multiple isoforms) and can open a discussion on which sequence to pick. First, be sure the sequence is from *Drosophila melanogaster*, then look for **Full Protein**. If **Full Protein** does not exist, pick the best choice (first isoform or largest size). Select the entry by clicking on the title. You will be brought to the flat file or submission entry for that sequence.

d. Flat files contain a lot of useful information but not in the most accessible format. Some translation for the students is necessary. You want to scroll down to the first literature reference associated with this sequence. If it is a primary article specific for the gene of study, it is a good choice. However, if the first reference is for a whole genome project, it is not specific for your gene and you may have a gene prediction.

For example, I searched for *couch potato Drosophila* and received 364 entries with multiple isoforms. When I amended the search to *couch potato Drosophila full*, I only received one entry. In that entry, the first reference to the primary literature in the flat file was:

```
AUTHORS     Bellen,H.J., Kooyer,S., D'Evelyn,D. and Pearlman,J.

TITLE       The Drosophila couch potato protein is expressed in nuclei of periperal
      neuronal precursors and shows homology to RNA-binding proteins

JOURNAL     Genes Dev. 6 (11), 2125-2136 (1992)
```

The title to this article provides some meaningful information. The gene name is mentioned and the title indicates that expression studies were performed. The date indicates that it is pre-genome sequencing projects (before 2000) and it has less than 10 authors. This entry indicates that this is a good sequence to use because it is based on the characterization of an individual gene.

Entries to avoid are the following:

```
AUTHORS     Adams,M.D., et al.(almost 100 authors),

TITLE       The genome sequence of Drosophila melanogaster

JOURNAL     Science 287 (5461), 2185-2195 (2000)
```

Such an entry indicates no experimental work was performed on the individual gene, but that it is part of a bulk download of genomic sequence. If this reference is the only one associated with the sequence, then the sequence is not the best choice and may be a prediction or a variant. Not every gene in GenBank has the same level of experimental data to support a predicted role.

e. Now that the most meaningful and best-supported sequence is selected, go to the top of the flat file, and select FASTA (under the gene name in the title). Hopefully, you see:

RecName: Full=Protein couch potato

UniProtKB/Swiss-Prot: Q01617.3

GenPept Graphics

>gi|48429205|sp|Q01617.3|CPO_DROME RecName: Full=Protein couch potato

MVKIANYQDLLGSHHQLLIAATAAAAAAAAAEPQLQLQHLLPAAPTTPAVISNPINSIGPINQISSSSHP
SNNNQQAVFEKAITISSIAIKRRPTLPQTPASAPQVLSPSPKRQCAAAVSVLPVTVPVPVPVSVPLPVSV
PVPVSVKGHPISHTHQIAHTHQISHSHPISHPHHHQLSFAHPTQFAAAVAAHHQQQQQQQAQQQQQAVQQ
QQQQAVQQQQVAYAVAASPQLQQQQQQQQQHRLAQFNQAAAAALLNQHLQQQHQAQQQQHQAQQQSLAHYG
GYQLHRYAPQQQQQHILLSSGSSSSKHNSNNNSNTSAGAASAAVPIATSVAAVPTTGGSLPDSPAHESHS
HESNSATASAPTTPSPAGSVTSAAPTATATAAAAGSAAATAAATGTPATSAVSDSNNNLNSSSSSNSNSN
AIMENQMALAPLGLSQSMDSVNTASNEEEVRTLFVSGLPMDAKPRELYLLFRAYEGYEGSLLKVTSKNGK
TASPVGFVTFHTRAGAEAAKQDLQGVRFDPDMPQTIRLEFAKSNTKVSKPKPQPNTATTASHPALMHPLT
GHLGGPFFPGGPELWHHPLAYSAAAAAELPGAAALQHATLVHPALHPQVPTQMTMPPHHQTTAIHPGAAM
AHMAAAAAAAAAGGGGGAATAAAAPQSAAATAAAAAAASHHHYLSSPALASPAGSTNNASHPGNPQIAAN
APCSTLFVANLGQFVSEHELKEVFSSHGNSNWLKLLHQ

The protein is in the FASTA format appropriate for conducting BLAST analysis. The sequence can be copied into a simple text program and saved.

f. Go to http://www.beanbeetles.org/genome/blast/beetleblast/beetleblast.php

Paste the FASTA file into the search box. Select program tblastn (to use your protein sequence to search the translated nucleotide database) and select bean beetle database. Then select basic search. Output should look like:

TBLASTN 2.2.27+

Reference:
Stephen F. Altschul, Thomas L. Madden, Alejandro A. Schäffer,
Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman (1997),
“Gapped BLAST and PSI-BLAST: a new generation of protein database
search programs”, Nucleic Acids Res. 25:3389-3402.

Database: ./db/longContigs.fasta
85,859 sequences; 315,317,553 total letters

```
Query= gi|48429205|sp|Q01617.3|CPO_DROME RecName: Full=Protein couch potato
Length=738

                                                                      Score     E
Sequences producing significant alignments:                          (Bits)  Value

  scaffold283865                                                      60.8    2e-08
  scaffold268507                                                      57.8    2e-07
  scaffold225587                                                      39.3    0.087
  scaffold50631                                                       33.5    6.6

> gi|48429205|sp|Q01617.3|CPO_DROME on scaffold283865
Length=2108

 Score = 60.8 bits (146),  Expect = 2e-08, Method: Compositional matrix adjust.
 Identities = 31/41 (76%), Positives = 32/41 (78%), Gaps = 1/41 (2%)
 Frame = -1

Query  522   MPQTIRLEFAKSNTKVSKPKPQPNTATTASHPALMHPLTGH  562
             MPQTIRLEFAKSNTKVSKPK Q   A   +HP LMHPLTG
Sbjct  1049  MPQTIRLEFAKSNTKVSKPKQQATNAAN-THPTLMHPLTGR  930


> gi|48429205|sp|Q01617.3|CPO_DROME on scaffold268507
Length=14563

 Score = 57.8 bits (138),  Expect = 2e-07, Method: Compositional matrix adjust.
 Identities = 28/53 (53%), Positives = 36/53 (68%), Gaps = 0/53 (0%)
 Frame = +3

Query  684   GSTNNASHPGNPQIAANAPCSTLFVANLGQFVSEHELKEVFSSHGNSNWLKLL  736
             GS+++    G     +N PCSTLFVANLGQFVSEHELKE+F+ + +   L  L
Sbjct  6555  GSSSSQPGVGGGMGVSNHPCSTLFVANLGQFVSEHELKEIFARYESRTVLMFL  6713


> gi|48429205|sp|Q01617.3|CPO_DROME on scaffold225587
Length=5685

 Score = 39.3 bits (90),  Expect = 0.087, Method: Compositional matrix adjust.
```

```
 Identities = 18/19 (95%), Positives = 19/19 (100%), Gaps = 0/19 (0%)
 Frame = +1

Query  475   EGYEGSLLKVTSKNGKTAS  493
             +GYEGSLLKVTSKNGKTAS
Sbjct  4486  QGYEGSLLKVTSKNGKTAS  4542


> gi|48429205|sp|Q01617.3|CPO_DROME on scaffold50631
Length=9815

 Score = 33.5 bits (75),  Expect = 6.6, Method: Compositional matrix adjust.
 Identities = 14/35 (40%), Positives = 23/35 (66%), Gaps = 0/35 (0%)
 Frame = -1

Query  423   MENQMALAPLGLSQSMDSVNTASNEEEVRTLFVSG  457
             +E Q  L  LG+ +  +S+ T SNE+ ++ LF+SG
Sbjct  1406  LEKQFILLSLGIPREQESLCTLSNEQYLQVLFISG  1302



Lambda      K        H        a         alpha
   0.316    0.129    0.388    0.792     4.96

Gapped
Lambda      K        H        a         alpha    sigma
   0.267   0.0410    0.140     1.90     42.6     43.6

Effective search space used: 58145294310


  Database: ./db/longContigs.fasta
    Posted date:  Mar 26, 2013  1:46 PM
  Number of letters in database: 315,317,553
  Number of sequences in database:  85,859
```

```
Matrix: BLOSUM62
Gap Penalties: Existence: 11, Extension: 1
Neighboring words threshold: 13
Window for multiple hits: 40
```

g. Students will need to evaluate the quality of their hits based on sequence similarity, length and quality. (For example, four hits are found with couch potato but only two have expected values low enough for further consideration. A good cut off range is an e-values larger than 10-6). Click the sequences in the subject column and click submit to download complete scaffolds. These sequences include data beyond just the area of the hit. Students may want to annotate sequence region identified in the blast analysis, especially if the scaffold is large.

h. Use scaffold sequence to perform a *blastn* against the bean beetle genome. Can any regions of overlap be identified to extend the sequence?

i. Use scaffold sequence to perform a *blastx* against GenBank. This analysis can be used to confirm that the quality of the bean beetle sequence. If the sequence is a good candidate for a similar gene, the hits retrieved should list similar functions to the original fruit fly sequence. However, if the sequence was a weak hit, unrelated or unfamiliar function will be seen.

j. The sequence quality of the *Callosobruchus maculatus* genome is variable and there are gaps in the sequence. You may see tracks of Ns (bases that could not be determined). Individual sequence reads are small and it may not be possible to annotate the whole gene.

Mission, Review Process & Disclaimer

The Association for Biology Laboratory Education (ABLE) was founded in 1979 to promote information exchange among university and college educators actively concerned with teaching biology in a laboratory setting. The focus of ABLE is to improve the undergraduate biology laboratory experience by promoting the development and dissemination of interesting, innovative, and reliable laboratory exercises. For more information about ABLE, please visit **http://www.ableweb.org/**

Papers published in *Tested Studies for Laboratory Teaching: Peer-Reviewed Proceedings of the Conference of the Association for Biology Laboratory Education* are evaluated and selected by a committee prior to presentation at the conference, peer-reviewed by participants at the conference, and edited by members of the ABLE Editorial Board.

Citing This Article

Ramesh, M. 2015. Mining the Genome of Callosobruchus maculatus (Bean Beetle) with a Little Help from Genetic Model Organisms. Article 49 in *Tested Studies for Laboratory Teaching,* Volume 36 (K. McMahon, Editor). Proceedings of the 36th Conference of the Association for Biology Laboratory Education (ABLE).
http://www.ableweb.org/volumes/vol-36/?art=49